Head CT Deep Learning Model for Early Stroke Identification Outperforms Human Experts

Bernardo Canedo Bizzo (✉ bbizzo@mgh.harvard.edu)

MGH & BWH Center for Clinical Data Science <https://orcid.org/0000-0002-9686-6751>

Romane Gauriau

MGH & BWH Center for Clinical Data Science

Donnella Comeau

MGH & BWH Center for Clinical Data Science

James Hillis

MGH & BWH Center for Clinical Data Science

Christopher Bridge

MGH & BWH Center for Clinical Data Science <https://orcid.org/0000-0002-2242-351X>

John Chin

MGH & BWH Center for Clinical Data Science

Jayashri Pawar

MGH & BWH Center for Clinical Data Science

Ali Pourvaziri

MGH & BWH Center for Clinical Data Science

Ivana Sesic

MGH & BWH Center for Clinical Data Science <https://orcid.org/0000-0002-7563-6834>

Elshaimaa Sharaf

MGH & BWH Center for Clinical Data Science

Jinjin Cao

MGH & BWH Center for Clinical Data Science

Flavia Noro

MGH & BWH Center for Clinical Data Science

Felipe Kitamura

Diagnósticos da América SA (Dasa) <https://orcid.org/0000-0002-9992-5630>

Keith Dreyer

Center for Clinical Data Science, Massachusetts General Brigham, Boston, MA

John Kalafut

GE Healthcare

Katherine Andriole

MGH & BWH Center for Clinical Data Science

Stuart Pomerantz

Department of Radiology, Massachusetts General Hospital

Ramon Gonzalez

Department of Radiology, Massachusetts General Hospital

Michael Lev

Massachusetts General Hospital <https://orcid.org/0000-0003-0236-7319>

Brief Communication

Keywords: deep learning model, head CT, Non-contrast head CT (NCCT), stroke

Posted Date: April 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-415673/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Non-contrast head CT (NCCT) is extremely insensitive for early (< 3-6hrs) acute infarct identification. We developed a deep learning model that detects and delineates early acute infarcts on NCCT, using diffusion MRI as ground truth (3,566 NCCT/MRI training pairs). The model substantially outperformed 3 expert neuroradiologists on a test set of 150 CT scans (sensitivity 96% model versus 61–66% experts); infarct volume estimates strongly correlated with those of diffusion MRI ($r^2 > 0.98$).

Main Body

Stroke is a significant public health issue, affecting approximately 13.7 million people annually and the second major cause of death and disability worldwide [1]. The management of acute ischemic stroke was revolutionized in 2018 with publication of the DAWN trial [2]. This study showed that the time window for safe and effective stroke treatment could be expanded from 6 to 24 hours post symptom onset, with appropriate patient selection using “advanced” CT or MR imaging to detect and estimate the volume of irreversibly ischemic infarct “core”. Specifically, stroke patients with intracranial vascular occlusions and “small” (<50mL) infarct cores, treated with endovascular thrombectomy (EVT), achieved a 49% rate of functional independence at 90-days, compared to only 13% with best medical therapy. A 50mL infarct volume threshold was chosen as an enrollment criterion to minimize the risk of intracranial hemorrhage (ICH) as a treatment complication. The resulting effect size of 36% (49-13%) remains among the highest of any stroke trial to date, especially considering the treatment window of up to one-full day after symptom onset, with a “number-needed-to-treat” of only 2.8. In DAWN and related late-window (6-24hr) treatment studies, infarct core volume was either determined using maximally efficient, ground truth, reference standard MR diffusion-weighted imaging (DWI) or approximated using CT perfusion imaging (CTP) [3-5]. Regardless of the imaging modality used for infarct core evaluation, however, all stroke clinical treatment trials have underscored the critical need for rapid, safe, highly sensitive and specific assessment, ideally minimizing cost, complexity, and technical variability [6, 7].

Only one major EVT clinical trial, MR CLEAN, which assessed treatment safety and efficacy in early stroke (<6hr), used non-contrast CT (NCCT) exclusively to both rule out ICH prior to enrollment and to estimate core infarct volume for subgroup analyses [7]. Unfortunately, both detection and volume estimation of early ischemic findings on NCCT – even by expert, subspecialty-certified neuroradiologists with decades of experience interpreting complex stroke scans – is significantly limited by the typically-subtle decreased X-ray attenuation and low contrast-to-noise ratios of acute infarcts. This poor conspicuity, attributable to the mildly reduced blood pool and early vasogenic edema of these developing lesions, is especially difficult to perceive in the first 3-6 hours after stroke onset, before blood brain barrier breakdown becomes well established [8, 9]. Even with interpretation by highly-trained readers using optimal image review display parameters, the sensitivity of NCCT for early (3-6hr) stroke detection has been reported to range as low as 43-71%, compared to 97% for DWI [8-10]. In a 2002 study comparing NCCT and DWI stroke detection within 3 hours of symptom onset, sensitivity for expert readers was 61% by CT and 91% by DWI; for novice readers, sensitivity was 46% by CT and 81% by DWI, with CT described as “little better than

flipping a coin” [11]. As such, DWI is considered the operational reference standard for rapid, accurate, emergency department assessment of brain tissue viability; it identifies regions of reduced water diffusivity attributable to cytotoxic edema, likely to be irreversibly infarcted even in the setting of early, robust restoration of critically ischemic cerebral blood flow [12].

In this study, we developed a deep learning model that detects, delineates, and estimates the volume of early acute infarction on NCCT, using diffusion MRI as ground truth (**Figures 1,2**). Our model, adapted from the U-Net architecture, takes an NCCT series as input and generates a segmentation mask of the early infarct changes, which is used to estimate infarct volume [21, *Methods*]. Model training relies on a large dataset of paired admission NCCT followed by ground truth DWI scans, acquired within a short time interval of one another. Infarct cores were segmented semi-automatically and segmentation masks for each pair were registered to the corresponding NCCT images (**Figure 2a**). Expert review of the test set, which included scans from two different vendors and 8 different scanner models (**Table, Methods**), was randomized with a different order of presentation for each radiologist. The experts recorded the presence or absence of acute infarct and categorized infarct volumes as >0-20mL, >20-50mL, or >50mL.

Our model significantly outperformed three expert neuroradiologists (mean 25-years’ experience, blinded to all other clinical/imaging data) for core infarct detection on an independent test set of 150 NCCT scans (sensitivity 96% model versus 61-66% experts, **Figure 1a**). Of these 150 scans, 90 were stroke-positive and 60 stroke-negative; for the stroke-positive scans, median time (a) from symptom-onset-to-NCCT was 3.7 hours (IQR 1.3-5.1 hours; 14 time-points unavailable) and (b) from NCCT-to-DWI was 28 minutes (IQR 22-36 min); median time from NCCT-to-DWI for stroke-negative scans was 5.9 hours (IQR 1.9-27.1).

Our model also approached the accuracy of ground truth DWI for core volume assessment ($r^2 > 0.98$, **Figure 1b**). Regarding the 50mL core infarct volume threshold used for patient selection in most late window clinical trials, our model correctly identified infarcts larger than 50mL with 97% (29/30) accuracy, compared to the three experts whose accuracy varied from 23% (7/30) to 47% (14/30; $p < 0.0001$). The experts failed to detect 7% (2/30) to 23% (7/30) of these large infarcts and categorized 17% (5/30) to 47% (14/30) as being <20mL. Our model also detected 100% (60/60) of strokes >20mL, of which the experts missed 18% (11/60) to 32% (19/60); indeed, as per the Bland-Altman plot (**Figure 1c**), the 95% confidence interval for mean DWI-NCCT core volume measurement was under ± 17 mL overall, across all volumes. Confusion matrices for infarct segmentation accuracy confirm superior model performance versus experts for estimating >0-20mL, >20-50mL, and >50mL volume thresholds (**Figure 1d**).

These results suggest that our model has the potential to obviate the need for more complex, costly, and time-consuming “advanced” CT and MR imaging (e.g., CTP, DWI) for safe, rapid assessment of infarct core - essential to patient selection for both early- and late time window stroke treatments such as EVT.

The performance of our model compared favorably to those of other, published AI models for NCCT acute stroke detection and delineation. This result is likely attributable to our large, accurately labeled training

set consisting of 3,566 NCCT / ground truth diffusion MRI pairs of early strokes (most <6hrs post-onset), for which DWI was obtained within 3-hours of admission CT for stroke-positive patients (median ≤ 50 min) and within 5-days for stroke-negative patients (median ≤ 19 hrs) (**Figure 2a**) [13-18]. It is noteworthy that both our training/validation and test sets contained predominantly small volume strokes (median DWI infarct volume <10mL and <30mL, respectively; see **Table, Methods**). Much of the existing work on automated detection and analysis of acute stroke focuses on three approaches: imaging features engineering, ischemic region segmentation, or biomarkers computation [13]. Although some of this literature reports high performance, few of these studies are focused on early ischemic findings and limitations include small and/or poorly labeled poorly annotated training datasets, as well as weaker “reference standard” ground truth (e.g., ground truth based on reader consensus or on less accurate, more highly variable modalities than MR-DWI, such as CTP) [3, 4, 13-18].

In one published model tested on 100 CT scans, for example (median 48-minutes after symptom onset, IQR 27-93 minutes), there was moderate correlation between algorithm-predicted NCCT and expert-contoured DWI infarct volumes ($r=0.76$, $r^2=0.58$), with the Bland-Altman plot 95% confidence interval for DWI-NCCT core volume measurement ranging from -59 to 80mL, versus -18 to 16mL for our model (**Figure 1c**) [14]. Recently, a model trained on NCCT/DWI pairs showed 0.76 accuracy for <9hr infarct detection [15]. For a different recently published model tested on 479 early and late window acute stroke CTs, there was modest correlation between NCCT predicted volumes and both CTP derived ($r=0.44$, $r^2=0.19$) and final-infarct ($r=0.52$, $r^2=0.27$) volumes [16]. Another recent model showed moderate performance in correlating automated NCCT Alberta Stroke Program Early CT Scores (i.e., “ASPECTS”, a 10-point scoring system for infarct size) with measured CTP ($r^2=0.58$) and DWI ($r^2=0.46$) core volumes [17]. Moreover, our algorithm’s accuracy is notably superior to that of the CTP derived infarct volume accuracies reported in the literature (e.g., Bland-Altman plot 95% confidence interval for mean CTP-DWI core volume measurement ranging from -59 to 55mL [18]).

Few medical artificial intelligence (AI) models to date have significantly outperformed human experts, and better-than-human detection and delineation of clinically important findings on CT or MRI cross sectional imaging has not previously been emphasized in the literature [19, 20]. In one study of a convolutional neural network (CNN) for malignant melanoma detection, compared to a group of 58 dermatologists with a broad range of experience including 30 experts, the “CNN missed fewer melanomas and misdiagnosed benign moles less often as malignant” [19]. In another AI imaging study, McKinney et al described a system for breast cancer screening mammography that outperformed US board certified radiologists “compliant with the requirements of the Mammography Quality Standards Act” [20]. There was a 5.7% reduction in false positives and a 9.4% reduction in false negatives with this system, which outperformed all human readers with an area under the receiver operating characteristic curve (AUC-ROC) of 0.740, reflecting an 11.5% improvement over the 0.625 AUC radiologist average. The authors concluded that AI has the potential to alleviate pressures on limited radiology staffing resources, as well as to discern “patterns and associations that are often imperceptible to humans”. Indeed, **Figure 2b** shows two head CT’s that were interpreted as negative for stroke by all three of our neuroradiology

experts, but correctly classified by our model as positive for early infarction (one of which had a large, >125mL infarct core).

In summary, we have developed a deep learning model that leverages the high sensitivity of DWI as ground truth to automate the detection, segmentation, and volume estimation of early ischemic changes on NCCT. Although DWI remains the reference standard for maximally sensitive, early infarct detection, MRI is a limited resource, not rapidly and routinely accessible in most acute care settings, such as community hospitals and rural urgent-care facilities, where only CT is likely to be available. Indeed, our deep learning platform might be especially beneficial to stroke patients in underserved areas, without 24/7 advanced imaging capability or off-hour radiologist staffing.

In conclusion, the accuracy of our AI model for non-contrast head CT early stroke *detection* and *volume estimation* (greater-than versus less-than 50mL) exceeds that of human experts, and approaches that of ground truth MR-DWI. If prospectively validated and confirmed to be generalizable across a variety of different CT-scanner platforms, manufacturers, and acquisition protocols at different institutions, this model has the potential to considerably reduce the need for more complex, costly, time-consuming and limited-availability advanced CT and MR imaging techniques, for the safe rapid selection of patients for both early- and late-time window highly effective stroke treatments such as endovascular thrombectomy.

Methods

This was a HIPAA-compliant retrospective study, with institutional review board approval and waived patient consent. The dataset was identified by searching the radiology exam archive of two large US academic medical centers (BB, SP) for non-contrast head CT (NCCT) scans for which patients also had MR diffusion-weighted imaging (DWI) scans acquired within the following 5 days. Brain MRI reports were screened using natural language processing to identify studies positive and negative for acute stroke. The time difference between the NCCT and MR-DWI imaging was limited to under 3 hours for stroke-positive scans, in order to capture infarct-related physiological changes on NCCT as close as possible to the MRI ground truth, and to under 5-days for stroke-negative scans, as restricted diffusion persists for several weeks following acute stroke. Parsing methods included keyword and sentence matching; all reports were manually reviewed by a trained radiologist (JP, JKC, BB). Scans were de-identified during image transfer using the Radiological Society of North America Clinical Trial Processor, with customized scripts to maintain relevant Digital Imaging and Communications in Medicine (DICOM) tags for series identification.

Brain MR-DWI and Apparent Diffusion Coefficient (ADC) sequences were considered ground truth for the presence or absence of acute infarction; axial DWI “b=1000” and ADC series with slice thickness $\geq 5\text{mm}$ were selected using a brain MRI series selection algorithm [22]. All images were reviewed by a trained radiologist (JKC, DC, BB, JP, AP, IS, JC) to ensure correct classification. Infarct segmentation was performed using established methodology, including a previously developed algorithm for mask generation [23]. The automated masks were reviewed by a trained radiologist along with the

corresponding MR-DWI/ADC series and radiology reports. Masks of positive scans were assigned a 5-point scale reflecting segmentation quality, where 4 reflected perfect overlap (i.e., 1:1 correspondence between DWI/ADC infarct and mask) and 0 reflected absent overlap. Only scans with quality grades 3-4 were used for model development; others were discarded or manually segmented by a trained radiologist (JKC, DC, AP, FN; Osirix MD v11.0.3). Segmentations were converted into Neuroimaging Informatics Technology Initiative (NIfTI) masks for machine learning model use.

The segmented DWI stroke-positive and negative scans were paired with the corresponding non-contrast head CT scans, obtained post-symptom onset but prior to DWI acquisition. Axial CT images with slice thickness $\leq 5\text{mm}$ and standard or soft kernel reconstructions, computed using routine iterative reconstruction or filtered back projection algorithms, were manually selected for model input (JKC, DC, JP, AP, JC, ES); in some cases, this resulted in several CT scans per patient (**Table**, 2nd row). Scans were excluded if they were non-diagnostic (e.g., severe metal or motion artifact). NCCT/DWI image pairs were spatially registered using a SimpleITK Python package (v1.2) with a multiscale affine transformation and mutual information loss. Registration results were assessed visually using a checkerboard display of the NCCT and co-registered DWI; failed or imprecisely registered images were excluded. The registration transformation was subsequently applied to the DWI acute infarct masks, to obtain a registered mask on the NCCT images.

The resulting dataset was randomly sampled to create the training, validation, and testing sets (**Table**); for the purposes of this study the test set was restricted to 150 patients. For the training and validation sets, all selected CT scans were retained for model building, even if there were multiple scans per patient, to maximize algorithm robustness at training and enhance algorithm evaluation at validation; all scans from the same patient were used exclusively for either training or validation, but not both. For the test set, only a single CT scan per patient was used; if more than one was available, the earliest (i.e., closest to admission) within the defined post-symptom onset timeframe was used, with 5mm-thick standard kernel reconstructed slices prioritized. The final test set included 60 stroke-negative control patients and 90 stroke-positive patients, distributed evenly with 30 NCCT/DWI pairs in each of the $>0\text{-}20\text{mL}$, $>20\text{-}50\text{mL}$, and $>50\text{mL}$ infarct volume categories. Only patients with strokes in the treatment-relevant middle cerebral artery vascular territory of the brain were selected for inclusion. The demographic characteristics of the training, validation, and test sets are shown in the **Table**.

We developed a neural network that can take as input 3D CT axial image stacks of varying numbers of slices, to generate segmentation masks as outputs. Pre-processing steps include resampling of the 3D input NCCT-image dataset and window/level pixel-value scaling. The resulting data becomes input to the neural network, which outputs both a classification result and a segmentation mask. Specifically: (1) each axial scan slice from the NCCT input is resized to a standardized 5mm thickness, then resampled to a 256x256 matrix size, for a maximum of 35 (inferior to superior) axial slices; (2) pixel intensities are clipped to window-width and center-level display range settings of 90 and 40 Hounsfield units, respectively, corresponding to the display parameters typically used clinically by neuroradiologists for workstation stroke CT image interpretation [8]; and (3) the resampled image pixel values are mapped

between 0 and 1; the binary masks, superimposed in the preprocessing step onto the original NCCT input slices, allows infarct volume estimation into $\leq 20\text{mL}$, $>20\text{-}50\text{mL}$, or $>50\text{mL}$ categories.

The network design extends the U-Net approach for biomedical image segmentation [21]. The 3D architecture is slightly modified, with an additional classification output (computed with a global max pooling from the segmentation output) that adds a classification component to the loss function, while maintaining segmentation and classification output consistency. Adding a classification component improved performance compared to using the Dice loss alone, as several very small stroke masks in our dataset could contribute disproportionately to lower the Dice score. The model was developed using Python 3.6 and Tensorflow 1.13.1. Moreover, although input image size is fixed in the axial in-plane dimensions, the framework can process 3D image volumes with varying numbers of slices. The architecture otherwise follows a classical U-Net design, with 6 down-sampling blocks (composed of 3×3 convolutions, batch normalization, and maximum pooling layers, followed by ReLU activation) and 6 up-sampling blocks; the main difference from a classical architecture is that the pooling operations are done at the slice level only, with shape (2, 2, 1), rather than between slices, which avoids unintended interpolation effects when the slice thickness is large. The neural network is optimized using a loss function that combines a differentiable dice loss (for segmentation) and a cross entropy loss (for study-level classification) as follows:

$L = L_{\text{dice}} + (1-\alpha) L_{\text{CE}}$, where L_{dice} is the dice loss, L_{CE} is the cross-entropy loss, and α is a constant ($0 \leq \alpha \leq 1$) reflecting the balance between segmentation and classification during training.

We applied geometrical and pixel intensity-based data augmentation techniques at the 3D volume level, which included a combination of in-slice rotations and translations, scaling, right-left flipping, and both Gaussian and Poisson random noise. At each epoch, each transformation was drawn with a probability of 0.5, and if applicable, the transformation parameters were randomly modified with a probability of 0.95. All transformations were applied in image space, prior to down-sampling, using linear interpolation. During training, CT series volume mini-batches were randomly selected for each epoch, without replacement. Because each mini-batch could contain a variable number of slices depending on the scan acquisition parameters, however, in order to both limit unnecessary memory and computational resource allocation, as well as to process standardized equally sized 3D volumes, the number of axial CT slices was fixed at 35, with generation of additional padded slices along the z-axis, as needed, if fewer than 35 were selected.

To control for data imbalance in our training set, we developed a standardized, batch sampling strategy. This included, *for each batch*: (1) selecting 8 stroke-positive and 4 stroke-negative scans, to ensure a fixed proportion of positive versus negative exams; and (2) selecting 7 scans acquired from General Electric (GE) CT platforms & 1 from Siemens platforms for stroke positive patients, and 2 from GE and 1 from Siemens for stroke negative patients, to reflect the manufacturer distribution of scanner platforms typically available for emergency department “stroke code” use at both institutions. Our datasets also included a broad range of small ($\leq 20\text{mL}$), intermediate ($\leq 50\text{mL}$), and large ($>50\text{mL}$) infarcts (**Table**).

Moreover, among stroke-positive scans, there was a large percentage of very small infarcts ($<1\text{ mL}$) in the training set ($455/1896=24\%$). Because signal-to-noise ratio, and hence CT conspicuity, of these tiny infarcts is likely to be poor - which could contribute to both decreased accuracy for stroke detection and increased error rate for small structure segmentation, impacting dice loss - we studied the effects on model performance of excluding infarcts smaller than 1 or 5 mL in our analyses (**Figure 1a**). Those results suggest that, for future clinical implementation, exclusion of infarcts smaller than 1 mL might provide an appropriate operating point on the ROC curve as a trade-off between optimizing both sensitivity and specificity for stroke *detection*.

Our neural network was trained using the Adam optimizer; network parameters were initialized with the uniform approach proposed by Glorot and Bengio [24]. The learning rate was reduced by a factor of 0.75 when the validation loss did not improve after 20 epochs. Our network trained for a maximum of 200 epochs, processed using NVIDIA 4 GPU Tesla V100 with 32Gb RAM, allowing batch sizes of twelve 3D volumes; training a single model took approximately 2.5 days. Such computationally demanding training was prohibitive for extensive hyperparameter search; approximately 400 different models were trained during the roughly 2-year development cycle. Hyperparameter search was performed manually with a grid search approach; the following parameters were tuned: learning rate, loss weights, batch sampling strategy (random uniform, positive/negative sampling, manufacturer sampling), exclusion/inclusion of infarcts ($< 1\text{ mL}$, $<5\text{ mL}$), and size of the first convolutional layer. After curation and data cleaning, several models were refined, and some hyper-parameters were adjusted. Hyper-parameter tuning was performed on the validation and training sets exclusively. Next, a small set of models were selected according to pre-defined performance metrics, including but not limited to Dice scores for the segmentation masks, ROC-AUC, sensitivity/specificity for stroke detection and volume estimation at the $>0\text{-}20\text{ mL}$, $>20\text{-}50\text{ mL}$, and $>50\text{ mL}$ segmented thresholds. These models were presented to a panel of several experienced radiologists (DC, BB, JKC), blinded to the specific model parameters, but with the performance metrics and a random, representative sample of results available for review for each model. The experts ranked these models and provided justification for their ratings; majority voting was used to select the final model to use for test set comparison to three, independent, expert neuroradiologists (ML, GG, SP) (**Figure 1a**).

For model metrics, 95% confidence intervals were computed using either the simple asymptotic method (for classification metrics) or bootstrapping technique (for continuous values, bootstrap size 500). Bland-Altman plot analysis was performed with MedCalc software (MedCalc for Windows, v19.8 / 2021, Ostend, Belgium). Python (v3.7) with NumPy package (v1.2) was used for all other statistical calculations, including but not limited to ROC curve analyses and linear regression. A $p<0.05$ level of confidence was considered statistically significant.

Declarations

Data availability

The training, validation, and test datasets generated for this study are protected patient information. Some data may be available for research purposes from the corresponding author upon reasonable request.

Code availability

The code base for the deep-learning framework makes use of proprietary components and we are unable to publicly release the full code base. However, all experiments and implementation details are described in sufficient detail in the Methods to enable independent replication with non-proprietary libraries.

References

1. Lindsay MP, *et al.* World Stroke Organization (WSO): Global Stroke Fact Sheet 2019. *Int J Stroke* **14**, 806-817 (2019).
2. Nogueira RG, *et al.*; DAWN Trial Investigators. Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct. *N Engl J Med.* **378**, 11-21 (2018).
3. Leslie-Mazwi TM, *et al.* Endovascular Stroke Treatment Outcomes After Patient Selection Based on Magnetic Resonance Imaging and Clinical Criteria. *JAMA Neurol.* **73**, 43-9 (2016).
4. Campbell BCV, *et al.*; HERMES collaborators. Penumbral imaging and functional outcome in patients with anterior circulation ischaemic stroke treated with endovascular thrombectomy versus medical therapy: a meta-analysis of individual patient-level data. *Lancet Neurol.* **18**, 46-55 (2019).
5. Nogueira RG, *et al.*; Trevo Registry and DAWN Trial Investigators. Stroke Imaging Selection Modality and Endovascular Therapy Outcomes in the Early and Extended Time Windows. *Stroke.* **52**, 491-497 (2021).
6. Kim BJ, *et al.* Endovascular Treatment After Stroke Due to Large Vessel Occlusion for Patients Presenting Very Late from Time Last Known Well. *JAMA Neurol.* **78**, 21–29 (2021).
7. Berkhemer OA, *et al.* A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. *N Engl J Med.* **372**, 11–20 (2015).
8. Lev MH, *et al.* Acute stroke: improved nonenhanced CT detection—benefits of soft-copy interpretation by using variable window width and center level settings. *Radiology* **213**, 150-5 (1999).
9. Camargo EC, *et al.* Acute brain infarct: detection and delineation with CT angiographic source images versus nonenhanced CT scans. *Radiology* **244**, 541-8 (2007).
10. Mullins, *et al.* CT and Conventional and Diffusion-Weighted MR Imaging in Acute Stroke: Study in 691 Patients at Presentation to the Emergency Department. *Radiology* **224**, 353–60 (2002).
11. Fiebach JB, *et al.* CT and diffusion-weighted MR imaging in randomized order: diffusion-weighted imaging results in higher accuracy and lower interrater variability in the diagnosis of hyperacute ischemic stroke. *Stroke* **33**, 2206-10 (2002).
12. Heiss WD, *et al.* Probability of cortical infarction predicted by flumazenil binding and diffusion-weighted imaging signal intensity: a comparative positron emission tomography/magnetic

- resonance imaging study in early ischemic stroke. *Stroke* **35**, 1892-8 (2004).
13. Mikhail P, Le MGD, Mair G. Computational Image Analysis of Nonenhanced Computed Tomography for Acute Ischaemic Stroke: A Systematic Review. *J Stroke Cerebrovasc Dis.* **29**, 104715 (2020).
 14. Qiu W, *et al.* Machine Learning for Detecting Early Infarction in Acute Stroke with Non-Contrast-enhanced CT. *Radiology* **294**, 638-644 (2020).
 15. Pan J, *et al.* Detecting the Early Infarct Core on Non-Contrast CT Images with a Deep Learning Residual Network. *J Stroke Cerebrovasc Dis.* **30**, 105752 (2021).
 16. Bouslama M, *et al.* Noncontrast Computed Tomography e-Stroke Infarct Volume Is Similar to RAPID Computed Tomography Perfusion in Estimating Postreperfusion Infarct Volumes. *Stroke* **52**, 634-641 (2021).
 17. Nagel S, *et al.* e-ASPECTS derived acute ischemic volumes on non-contrast-enhanced computed tomography images. *Int J Stroke.* **15**, 995-1001 (2020).
 18. Schaefer PW, *et al.* Limited reliability of computed tomographic perfusion acute infarct volume measurements compared with diffusion-weighted imaging in anterior circulation stroke. *Stroke.* **46**, 419-24 (2015).
 19. Haenssle HA, *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol.* **29**, 1836-1842 (2018).
 20. McKinney SM, *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89-94 (2020).

Methods References

21. Ronneberger O, Fischer P, and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Preprint at <https://arxiv.org/abs/1505.04597v1> (2015).
22. Gauriau R, *et al.* Using DICOM Metadata for Radiological Image Series Categorization: a Feasibility Study on Large Clinical Brain MRI Datasets. *J Digit Imaging.* **33**, 747-762 (2020).
23. Pedemonte S, *et al.* Detection and Delineation of Acute Cerebral Infarct on DWI Using Weakly Supervised Machine Learning. *Medical Image Computing and Computer Assisted Intervention (MICCAI)* **1107**, 81–88 (2018).
24. Glorot X and Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* **9**, 249-256 (2010).

Table

Table. *Dataset description with patient demographics and acquisition details*

(Legend: no.=number, Std=standard deviation, M/F=male/female, IQR=inter-quartile range,

mAs=milliampere-seconds, kVp=kilovoltage peak)

Patient demographics	Training	Validation	Test
No. patients (stroke positive / negative)	3566 (1896 / 1670)	133 (66 / 67)	150 (90 / 60)
No. NCCT scans	9528	338	150
Mean age (Std)	65 (17)	64 (17)	67 (17)
Gender: M / F	1834 / 1859	80 / 62	73 / 77
Infarct volume $\geq$ 50mL (no.)	334	17	30
Infarct volume $\geq$ 20mL (no.)	641	19	60
Median DWI infarct volume, mL	9.4	5.2	29.6
Acquisition			
Mean time from NCCT-to-DWI: stroke positive / stroke negative	50min / 16hrs	50min / 9hrs	35min / 19hrs
No. CT scans per vendor: GE Healthcare / Siemens (No. different CT models per vendor)	2509 / 1184	88 / 54	110 / 40 (5 / 3)
Range of years from which NCCT/DWI scans obtained	2001 - 2019	2002 - 2019	2002 - 2019
Mean X-ray scanning current	225 mA	217 mA	220 mA
Mean X-ray scanning voltage	119 kVp	120 kVp	122 kVp

Figures

Figure 1

Model performance for infarct detection (a, ROC curve) and delineation (b, scatterplot; c, Bland-Altman plot; d, confusion matrices), based on DWI ground truth, compared to three human experts. (a) Model AUC was 0.95; sensitivity/specificity were 0.96/0.72 at a 0mL-threshold operating point for infarct detection, 0.82/0.92 at a 1mL-threshold, and 0.78/0.98 at a 5mL-threshold for infarct detection, compared to mean reader sensitivity/specificity of 0.64/0.91. (b) Model infarct volume estimates strongly correlated with those of DWI ground truth ($r^2 > 0.98$). As per the Bland-Altman plot (c), the model had excellent performance for distinguishing infarcts smaller versus larger than 50mL (95%CI<+17mL), the volume threshold used for patient selection in major late window stroke treatment trials. Expert interrater Cohen's kappa values ranged from 0.42-0.48, suggesting significant variability compared to the model, confirmed by the confusion matrices for volume segmentation (d, mean study-counts-per-category and ranges shown for the 3-experts; calculated at the model's 0mL-threshold for infarct detection).

Figure 2

(a) Schematic representation of model development steps from cohort creation and data preparation to model inference training. (b) Examples of infarcts not detected by each of the three expert neuroradiologists, but accurately detected and delineated by the model.